TRECVID-2014 Semantic Indexing task: Overview

Georges Quénot
Laboratoire d'Informatique de Grenoble

George Awad
Dakota Consulting, Inc

NIST
National Institute of Standards and Technology

Outline

- Task summary (Goals, Data, Run types, Metrics)
- Evaluation details
  - Inferred average precision
  - Participants
- Evaluation results
  - Hits per concept
  - Results per run
  - Results per concept
  - Significance tests
- Progress task results
- Localization subtask results
- Global Observations
- Issues

Semantic Indexing task

- Goal: Automatic assignment of semantic tags to video segments (shots)
- Secondary goals:
  - Encourage generic (scalable) methods for detector development.
  - Semantic annotation is important for filtering, categorization, searching and browsing.
- Participants submitted four types of runs:
  - Main run Includes results for 60 concepts, from which NIST evaluated 30
    - Localization run includes results for 10 pixel-wise localized concepts from the 60 evaluated concepts in main runs.
  - Progress run Includes results for 60 concept for 2 non-overlapping datasets, from which 1 dataset will be evaluated the next year.

Semantic Indexing task (data)

- SIN testing dataset
  - Main test set (IACC.2.B): 200 hours, with durations between 10 seconds and 6 minutes.
  - Progress test set (IACC.2.C): 200 hours and non overlapping from IACC.2
- SIN development dataset
  - (IACC.1.A, IACC.1.B, IACC.1.C & IACC.1.tv10.training): 800 hours, used from 2010 – 2012 with durations between 10 seconds to just longer than 3.5 minutes.
- Total shots:
  - Much more than in previous TRECVID years, no composite shots
  - Development: 549,434
  - Test: IACC.2.A (112,677), IACC.2.B (106,913), IACC.2.C (113,161)
- Common annotation for 346 concepts coordinated by LIG/LIF/Quaero from 2007-2013 made available.

Semantic Indexing task (Concepts)

- Selection of the 60 target concepts
  - Were drawn from 500 concepts chosen from the TRECVID “high level features” from 2005 to 2010 to favor cross-collection experiments Plus a selection of LSCOM concepts so that:
    - we end up with a number of generic-specific relations among them for promoting research on methods for indexing many concepts and using ontology relations between them.
    - we cover a number of potential subtasks, e.g. “persons” or “actions” (not really formalized)
  - It is also expected that these concepts will be useful for the content-based (instance) search task.
- Set of relations provided:
  - 427 “implies” relations, e.g. “Actor implies Person”
  - 559 “excludes” relations, e.g. “Daytime_Outdoor excludes Nighttime”

Semantic Indexing task (training types)

- Six training types were allowed:
  - A – used only IACC training data (42 runs)
  - B – used only non-IACC training data (0 runs)
  - C – used both IACC and non-IACC TRECVID (S&V and/or Broadcast news) training data (0 runs)
  - D – used both IACC and non-IACC non-TRECVID training data(29 runs)
  - E – used only training data collected automatically using only the concepts' name and definition (4 runs)
  - F – used only training data collected automatically using a query built manually from the concepts' name and definition (0 runs)

Semantic Indexing task (training types)

- Stricter interpretation of type A since 2014:
  - Use of components built using other training data (e.g. face detectors) was considered as acceptable as long as this was not for directly training the SIN target concepts (no sample directly annotated with SIN concepts used)
  - Generalization to the use of components like semantic descriptors trained on external data (e.g. ImageNet) was similar in principle but too close to type D
  - Partially re-trained deep networks are even closer
- Many runs submitted in 2013 and earlier as type A would be now requalified as type D with the new interpretation (not a problem)
- Results are now presented in a single table and plot for types A-D (the training type still appear un the run names)

30 Single concepts evaluated(1)

3 Airplane*
9 Basketball
10 Beach*
13 Bicycling
15 Boat_Ship*
17 Bridges*
19 Bus*
25 Chair*
27 Cheering
29 Classroom
31 Computers*
41 Demonstration_Or_Protest
59 Hand*
63 Highway
71 Instrumental_Musician*
80 Motorcycle*
83 News_Studio*
84 Nighttime
100 Running*
105 Singing*
112 Stadium
117 Telephones*
163 Baby*
261 Flags*
267 Forest*
274 George_Bush*
321 Lakes
359 Oceans
392 Quadruped*
434 Skier

-The 19 marked with "*" are a subset of those tested in 2013

10 Localization Concepts evaluated (2)

- [3] Airplane
- [15] Boat_ship
- [17] Bridges
- [19] Bus
- [25] Chair
- [59] Hand
- [80] Motorcycle
- [117] Telephones
- [261] Flags
- [392] Quadruped

Evaluation

- Task: Find shots that contain a certain concept, rank them according to confidence measure, submit the top 2000.
- The 30 evaluated single concepts were chosen after examining TRECVid 2013 60 evaluated concept scores across all runs and choosing the top 45 concepts with maximum score variation.
- Each feature assumed to be binary: absent or present for each master reference shot
- NIST sampled ranked pools and judged top results from all submissions
- Metrics: *inferred average precision per concept*
- Compared runs in terms of **mean** *inferred average precision* across the 30 concept results for main runs.

Inferred average precision (infAP)

- Developed* by Emine Yilmaz and Javed A. Aslam at Northeastern University
- Estimates average precision surprisingly well using a surprisingly small sample of judgments from the usual submission pools
- More features can be judged with same effort
- Increased sensitivity to lower ranks
- Experiments on previous TRECVID years feature submissions confirmed quality of the estimate in terms of actual scores and system ranking

* J.A. Aslam, V. Pavlu and E. Yilmaz, ***Statistical Method for System Evaluation Using Incomplete Judgments*** Proceedings of the 29th ACM SIGIR Conference, Seattle, 2006.

2014: mean extended Inferred average precision (xinfAP)

- 2 pools were created for each concept and sampled as:
  - Top pool (ranks 1-200) sampled at 100%
  - Bottom pool (ranks 201-2000) sampled at 11.1%

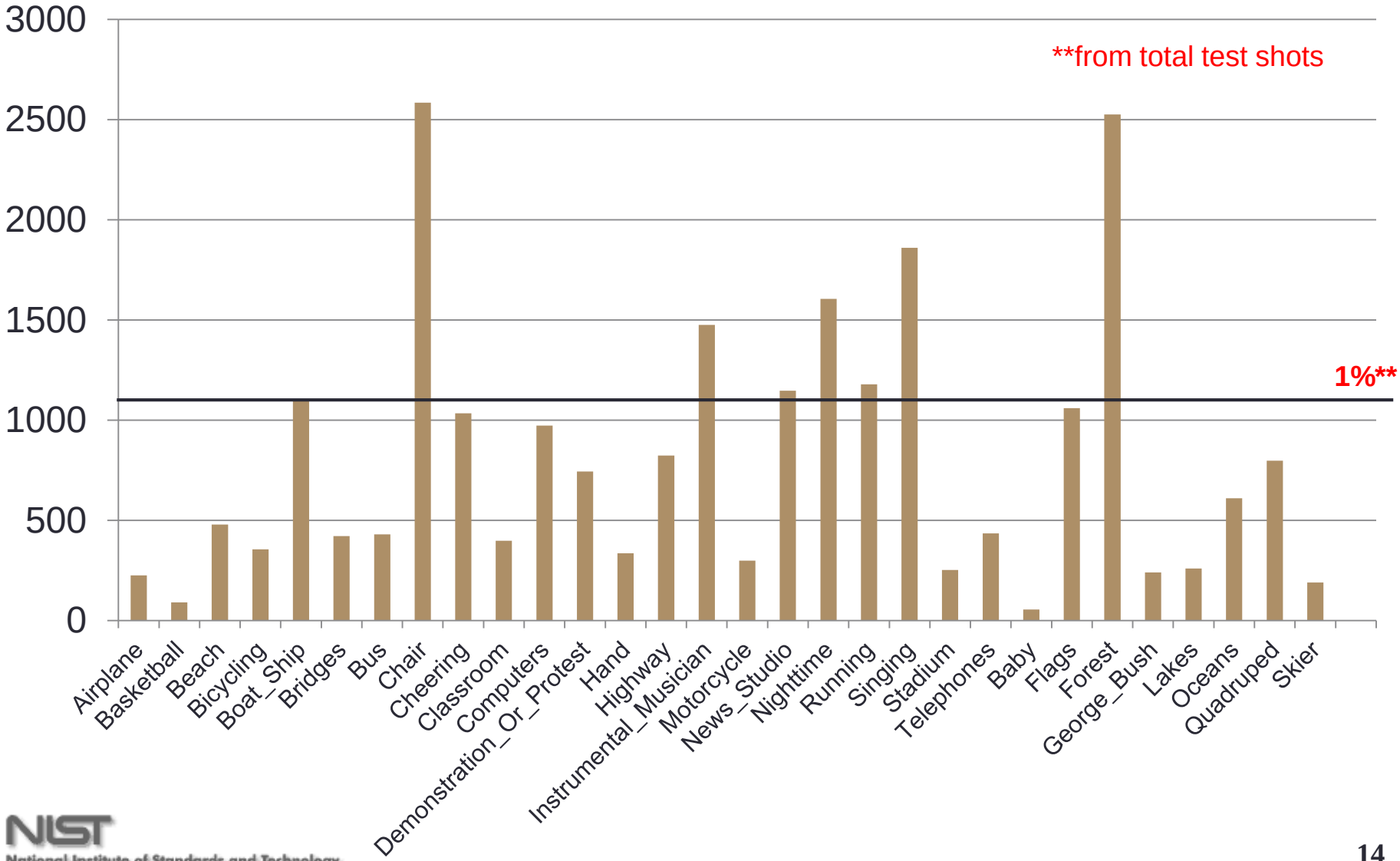
30 concepts
191,717 total judgments
12248 total hits
7938 Hits at ranks (1-100)
2869 Hits at ranks (101-200)
1441 Hits at ranks (201-2000)

- Judgment process: one assessor per concept, watched complete shot while listening to the audio.
- infAP was calculated using the judged and unjudged pool by sample_eval

2014 : 15 Finishers

CMU	Carnegie Mellon U.
CRCV_UCF	University of Central Florida
EURECOM	EURECOM - Multimedia Communications
FIU_UM	Florida International U., U. of Miami
Insightdcu	Insight Centre for Data Analytics
IRIM	CEA-LIST, ETIS, EURECOM, INRIA, LABRI, LIF, LIG, LIMSI, LIP6, LIRIS, LISTIC
ITI_CERTH	Information Technologies Institute, Centre for Research and Technology Hellas
LIG	Laboratoire d'Informatique de Grenoble
MediaMill	U. of Amsterdam
OrangeBJ	Orange Labs International Center Beijing
PicSOM	Aalto U.
PKUSZ_ELMT	Peking University Engineering Laboratory of 3D Media Technology
TokyoTech-Waseda	Tokyo Institute of Technology, Waseda University
UEC	U. of Electro-Communications
VIREO	City U. of Hong Kong

Inferred frequency of hits varies by concept



NIST
National Institute of Standards and Technology

Total true shots contributed uniquely by team

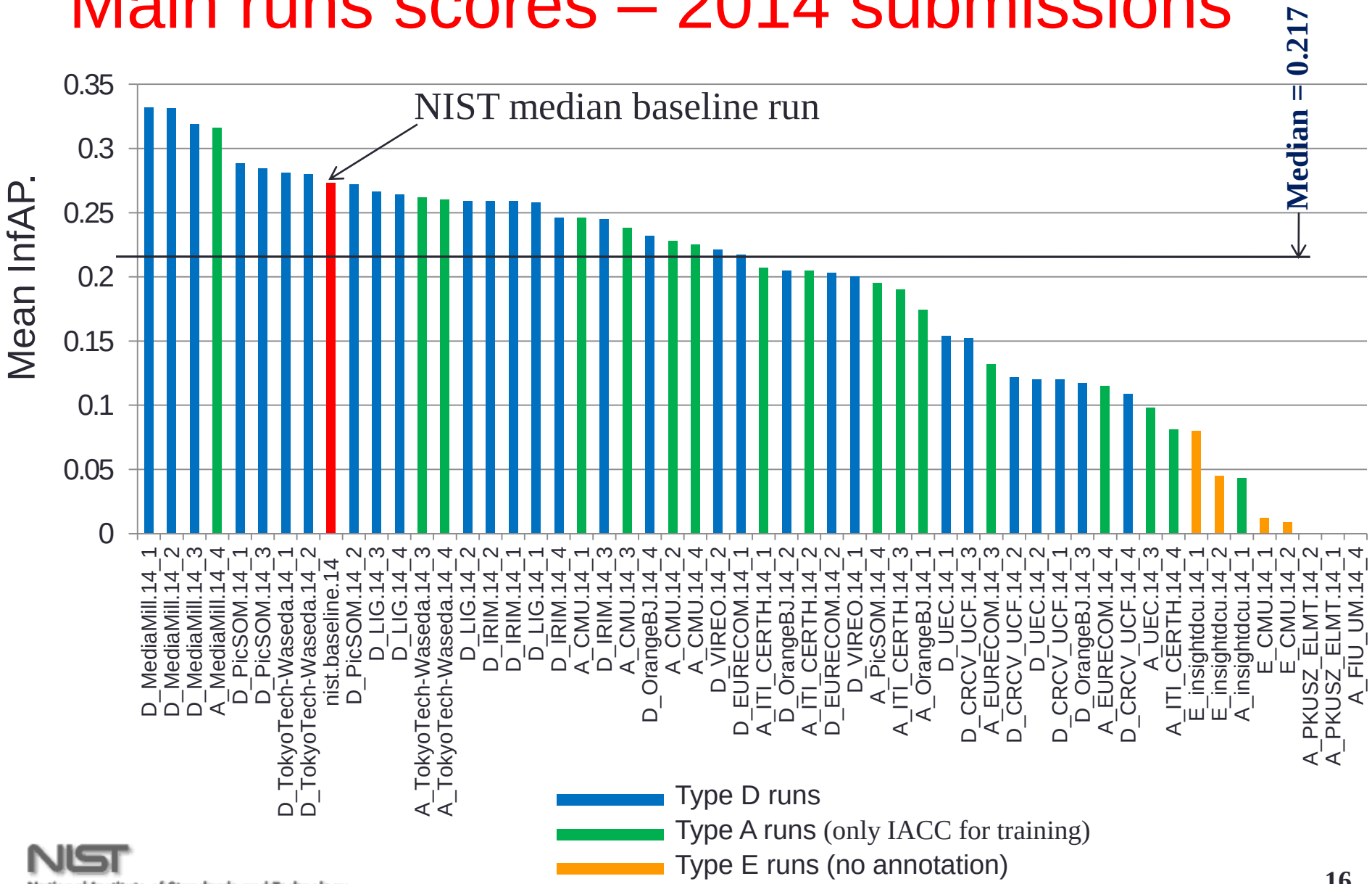
Team	No. of Shots	Team	No. of shots
Insightdcu	81	Mediamill	6
UEC	34	PKUSZ_ELMT	3
CMU	32	VIREO	2
EURECOM	24	LIG	1
OrangeBJ	22		
ITI_CERTH	19		
HFUT*	16		
FIU_UM	15		
NHKSTRL*	13		
NII*	13		
CRCV_UCF	11		
Picsom	11		
TokyoTech-Waseda	4		

Fewer unique shots compared to TV2013 & TV2012

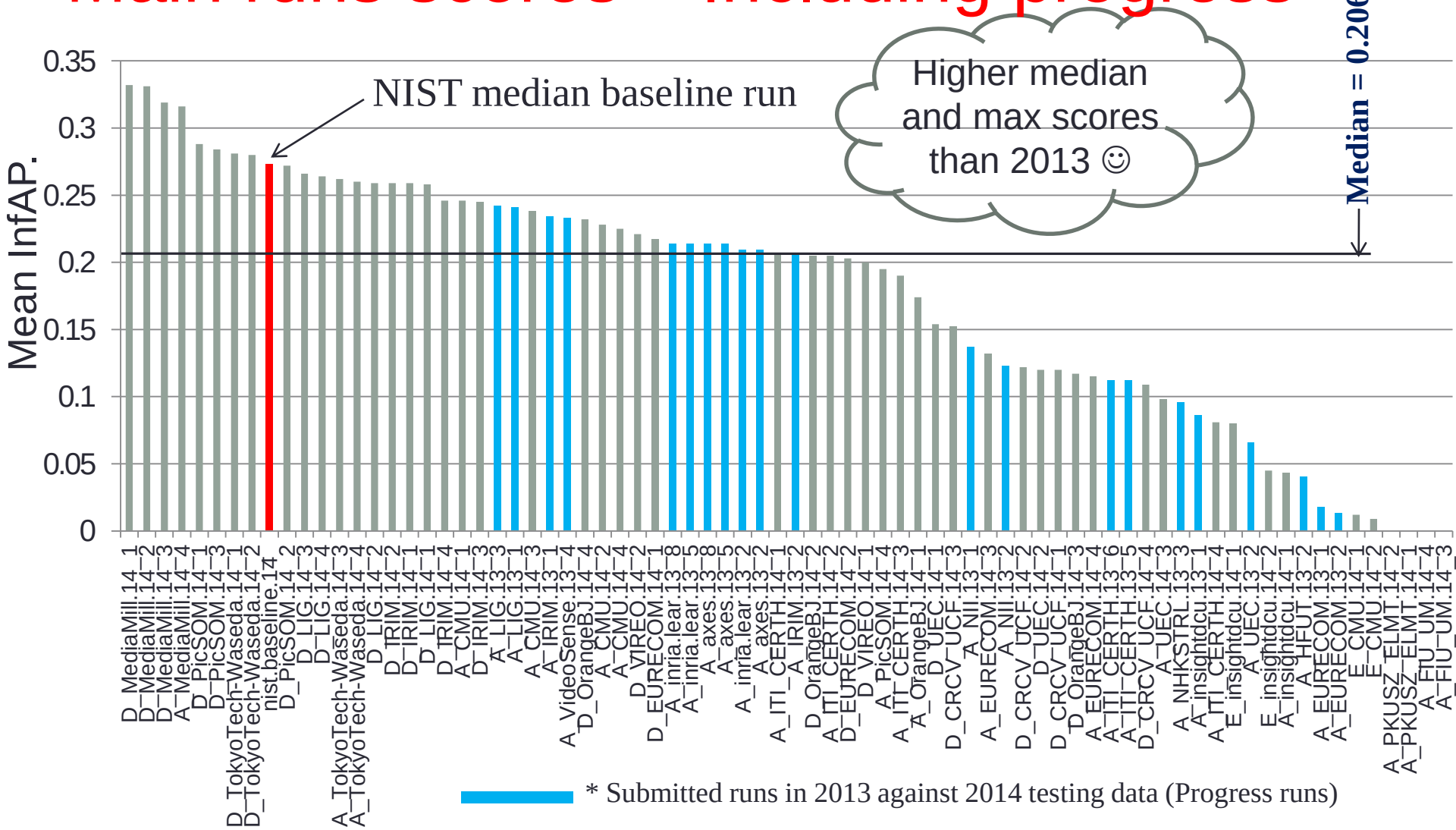
NIST National Institute of Standards and Technology

*shots submitted in 2013 in progress task

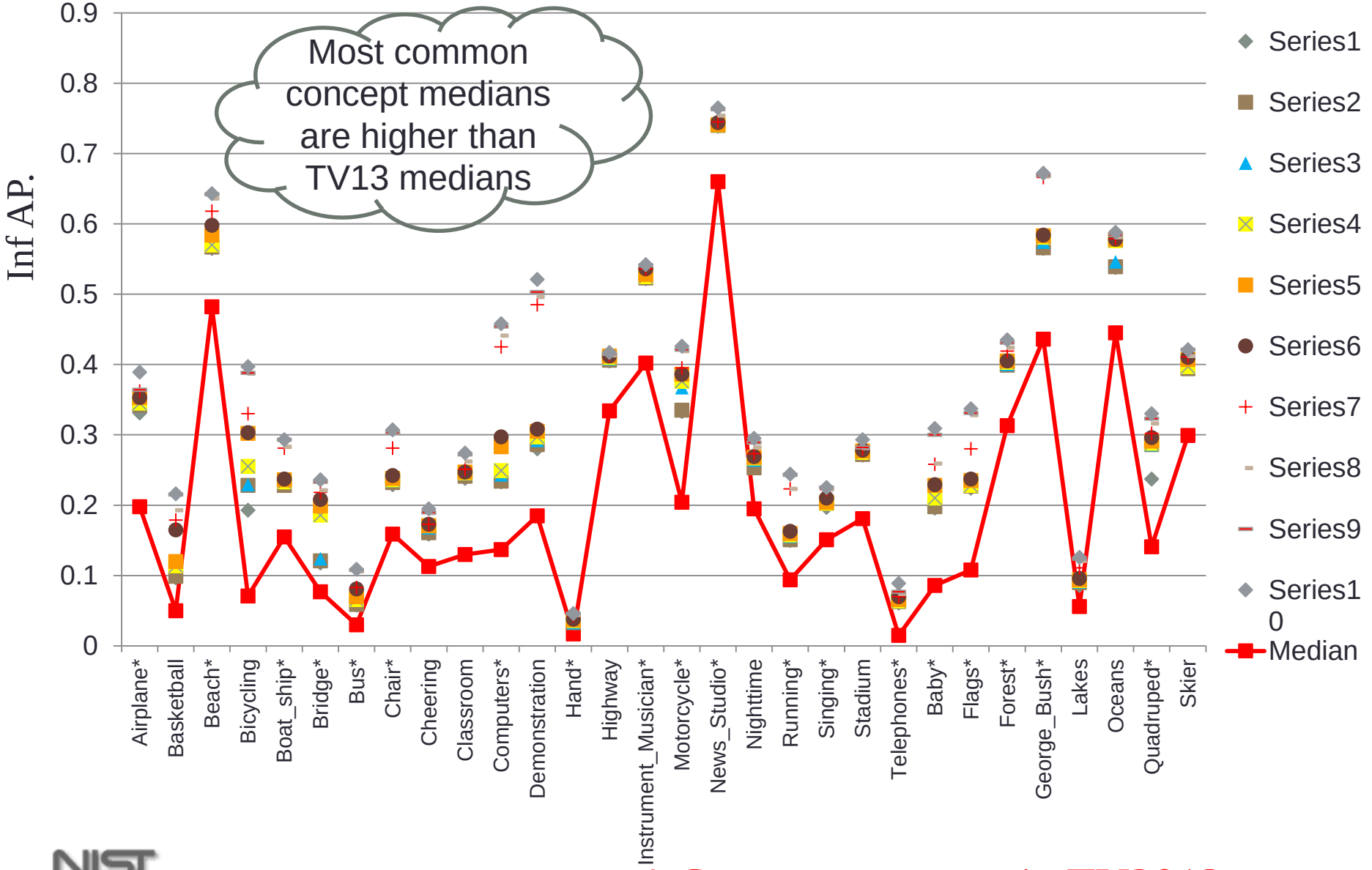
Main runs scores – 2014 submissions



Main runs scores – Including progress



Top 10 InfAP scores by concept (Main runs)



* Common concept in TV2013

Statistical significant differences among top 10 Main runs (using randomization test, p < 0.05)

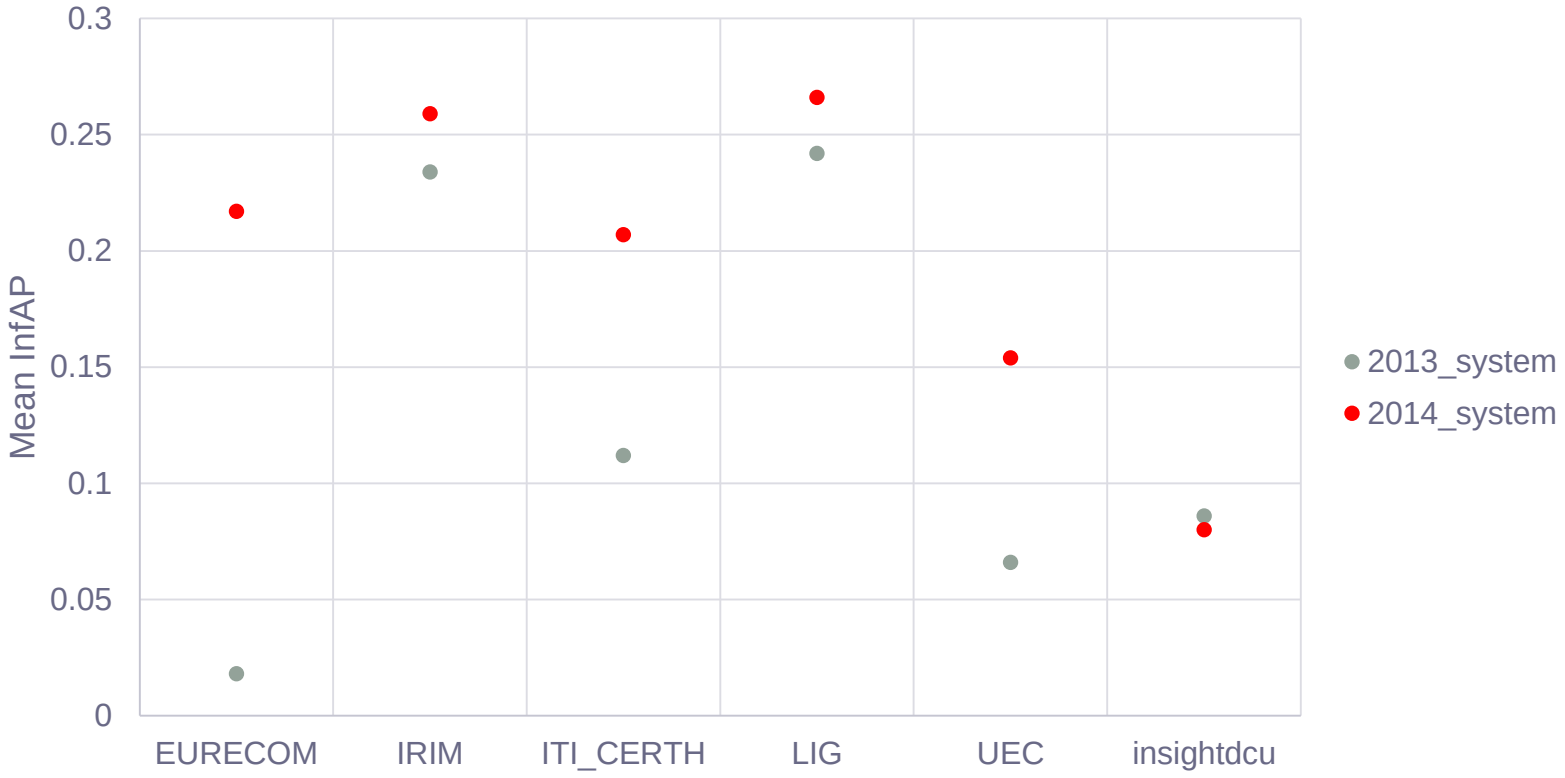
Run name	(mean infAP)
D_MediaMill.14_1	0.332
D_MediaMill.14_2	0.331
D_MediaMill.14_3	0.319
A_MediaMill.14_4	0.316
D_PicSOM.14_1	0.288
D_PicSOM.14_3	0.284
D_TokyoTech-Waseda.14_1	0.281
D_TokyoTech-Waseda.14_2	0.280
D_PicSOM.14_2	0.272
D_LIG.14_3	0.266

- D_MediaMill.14_1
  - D_MediaMill.14_3
    - D_TokyoTech-Waseda.14_2
    - D_TokyoTech-Waseda.14_1
      - D_LIG.14_3
    - D_PicSOM.14_1
      - D_PicSOM.14_3
        - D_PicSOM.14_2
        - D_LIG.14_3
  - D_MediaMill.14_4
    - D_TokyoTech-Waseda.14_2
    - D_TokyoTech-Waseda.14_1
      - D_LIG.14_3
    - D_PicSOM.14_1
      - D_PicSOM.14_3
        - D_PicSOM.14_2
        - D_LIG.14_3
- D_MediaMill.14_2
  - D_MediaMill.14_3
    - D_TokyoTech-Waseda.14_2
    - D_TokyoTech-Waseda.14_1
      - D_LIG.14_3
    - D_PicSOM.14_1
      - D_PicSOM.14_3
        - D_PicSOM.14_2
        - D_LIG.14_3
  - D_MediaMill.14_4
    - D_TokyoTech-Waseda.14_2
    - D_TokyoTech-Waseda.14_1
      - D_LIG.14_3
    - D_PicSOM.14_1
      - D_PicSOM.14_3
        - D_PicSOM.14_2
        - D_LIG.14_3

Progress subtask

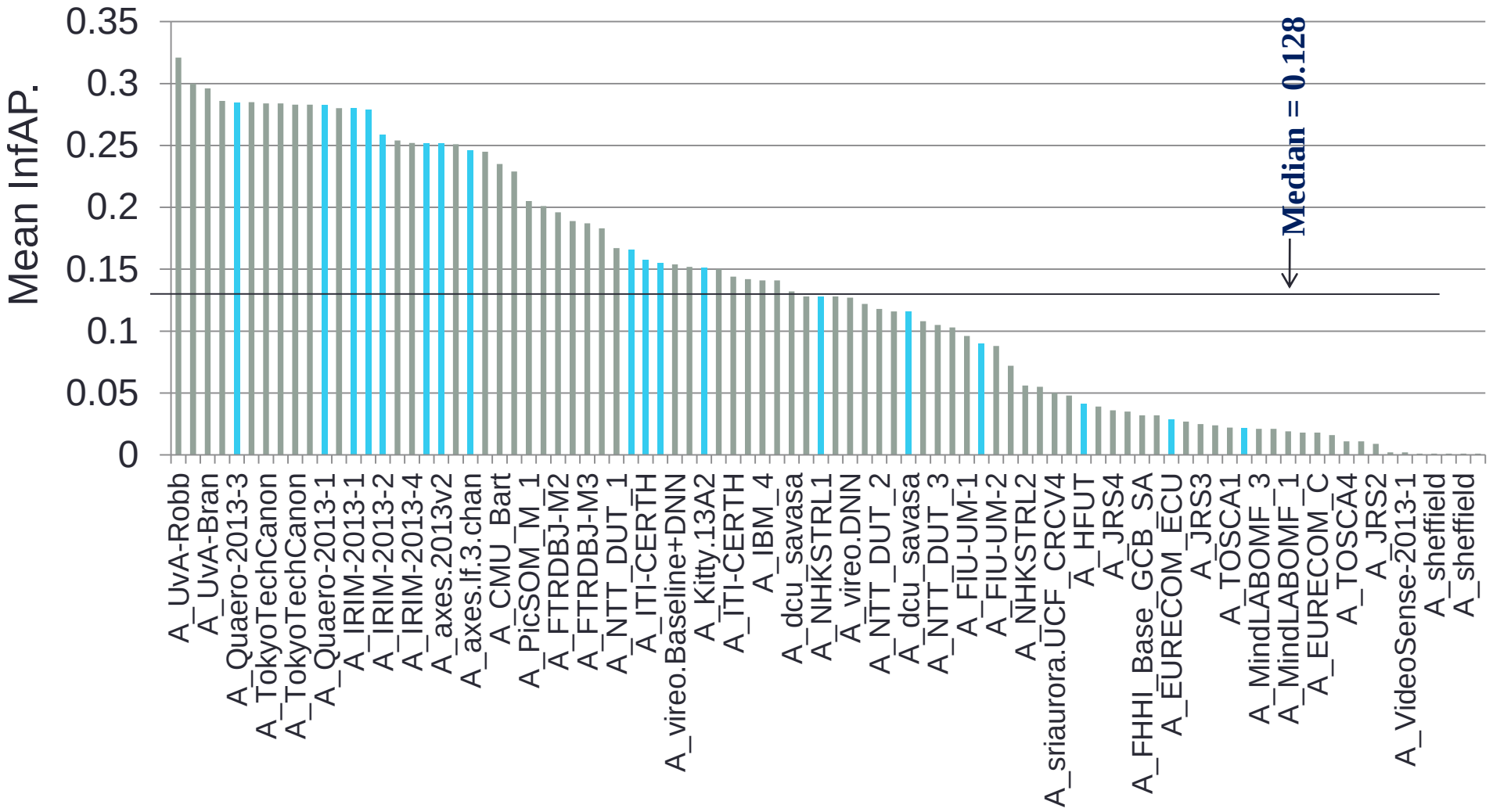
- Measuring progress of 2013 vs 2014 systems on IACC.2.B dataset.
- 2014 systems used same training data and annotations as in 2013.
- Total 6 teams submitted progress runs against IACC.2.B dataset.

Progress subtask: Comparing best runs in 2013 & 2014 by team

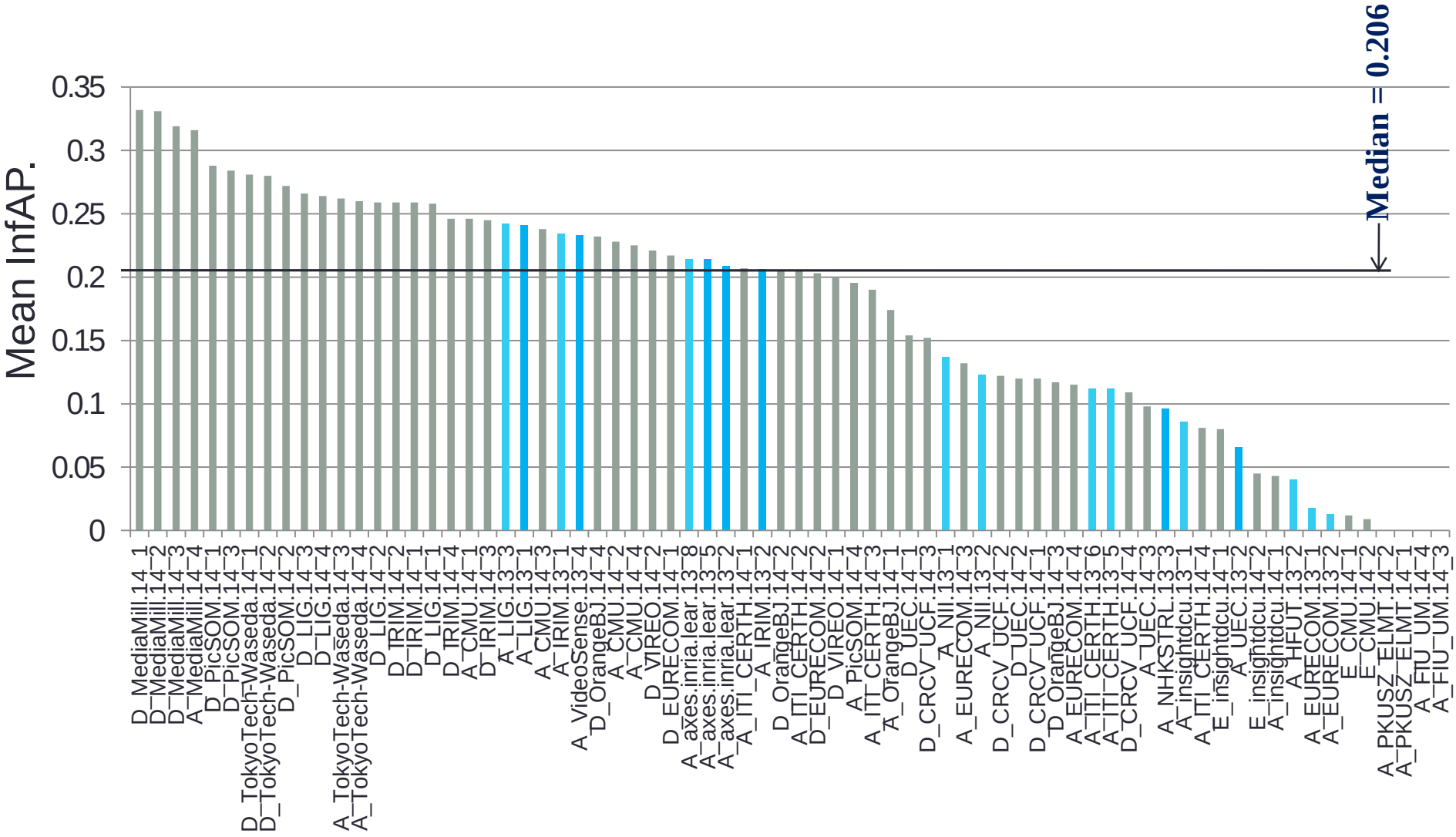


Randomization tests show that 2014 systems are better than 2013 systems (except for insightdcu)

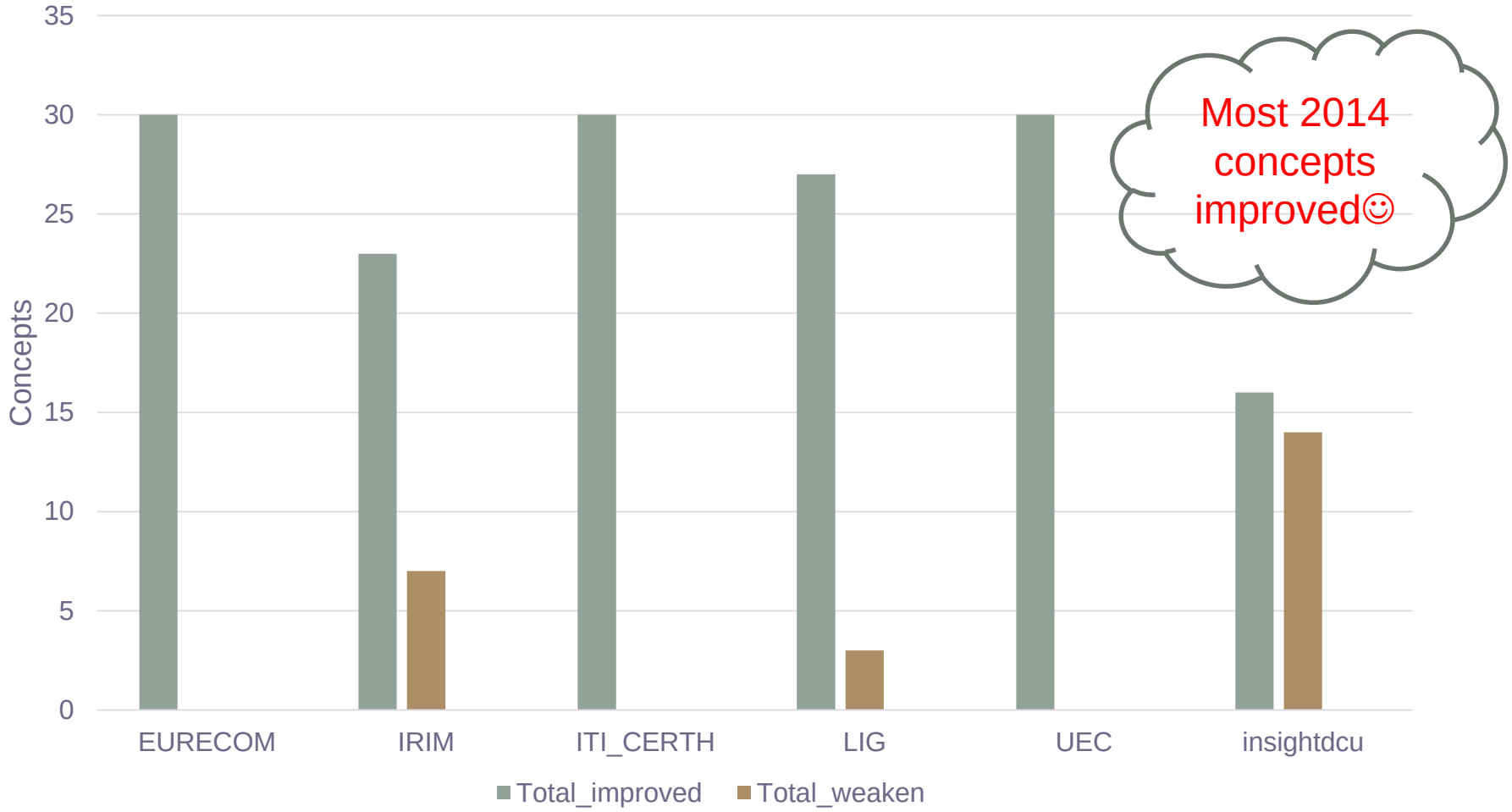
Main runs scores 2013



Main runs scores 2014



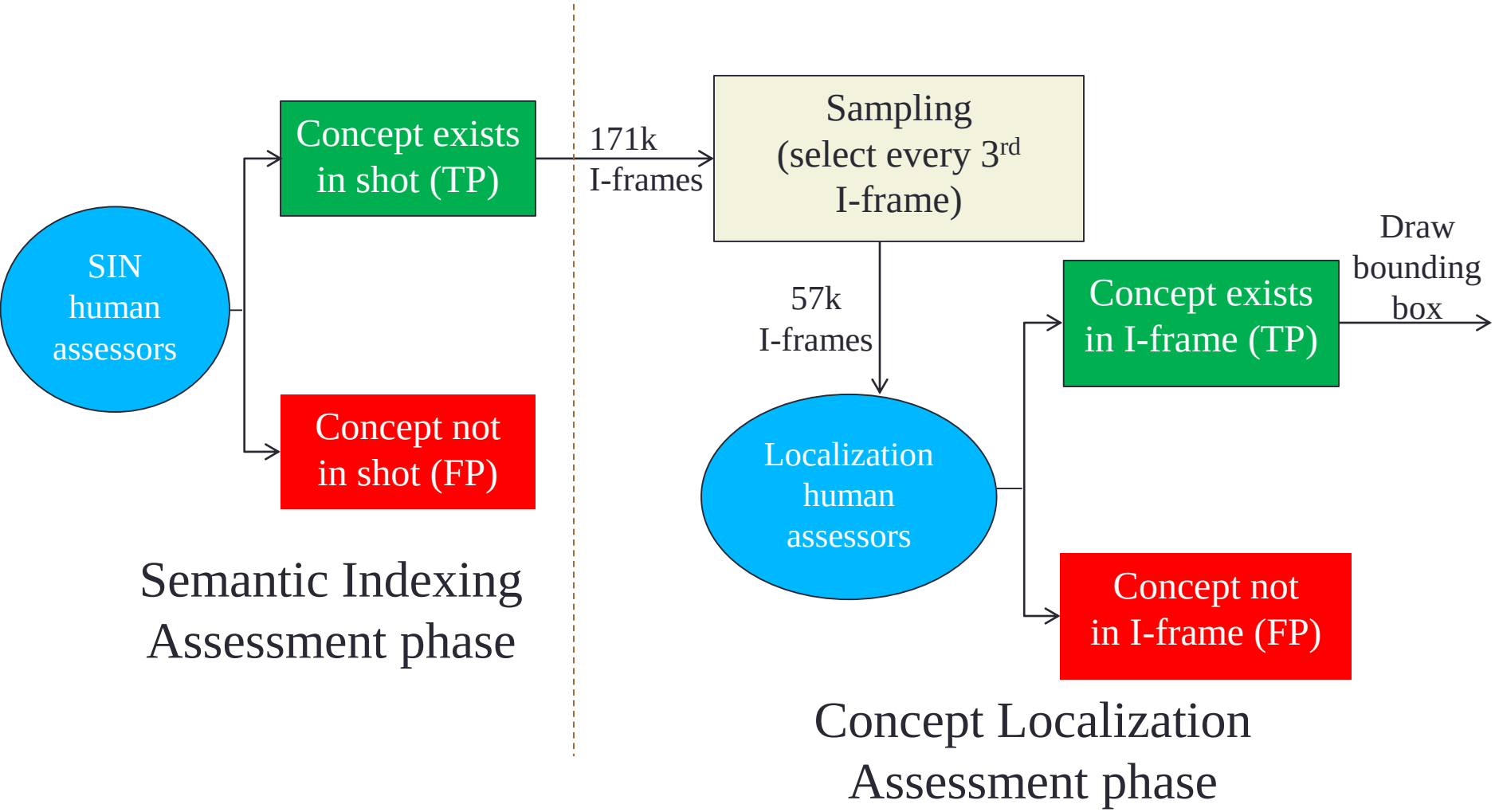
Progress subtask: Concepts improved vs weaken by team



Concept localization subtask

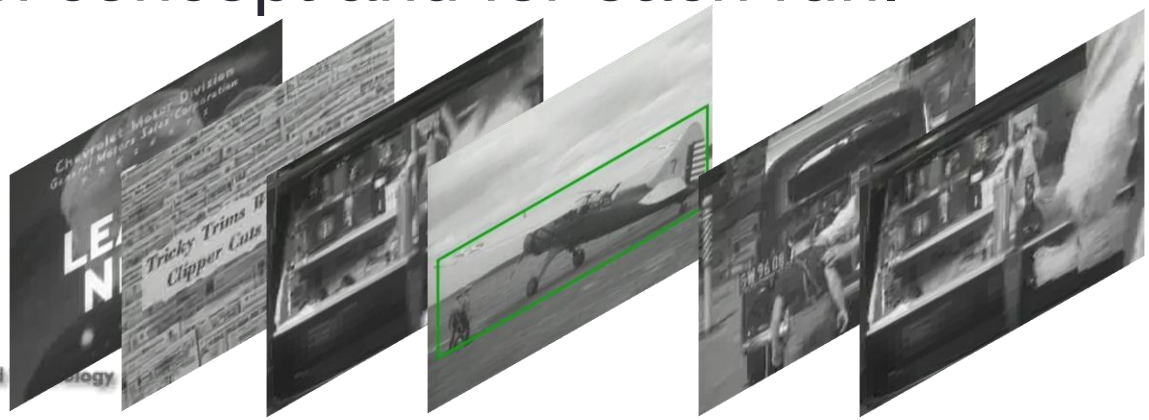
- Goal
  - Make concept detection more precise in time and space than current shot-level evaluation.
  - Encourage more reusable concept detectors design that is independent from the context.
- Task
  - For each of the 10 concepts
    - For each of the top 1000 main run shots in SIN run
      - For each I-Frame within the shot that contains the target, return
        - the x,y coordinates of the (UL,LR) vertices of a bounding rectangle containing all of the target concept and as little more as possible.
- Systems were allowed to submit more than 1 bounding box per I-frame but only the one with maximum fscore were judged.

NIST Evaluation framework



Evaluation metrics

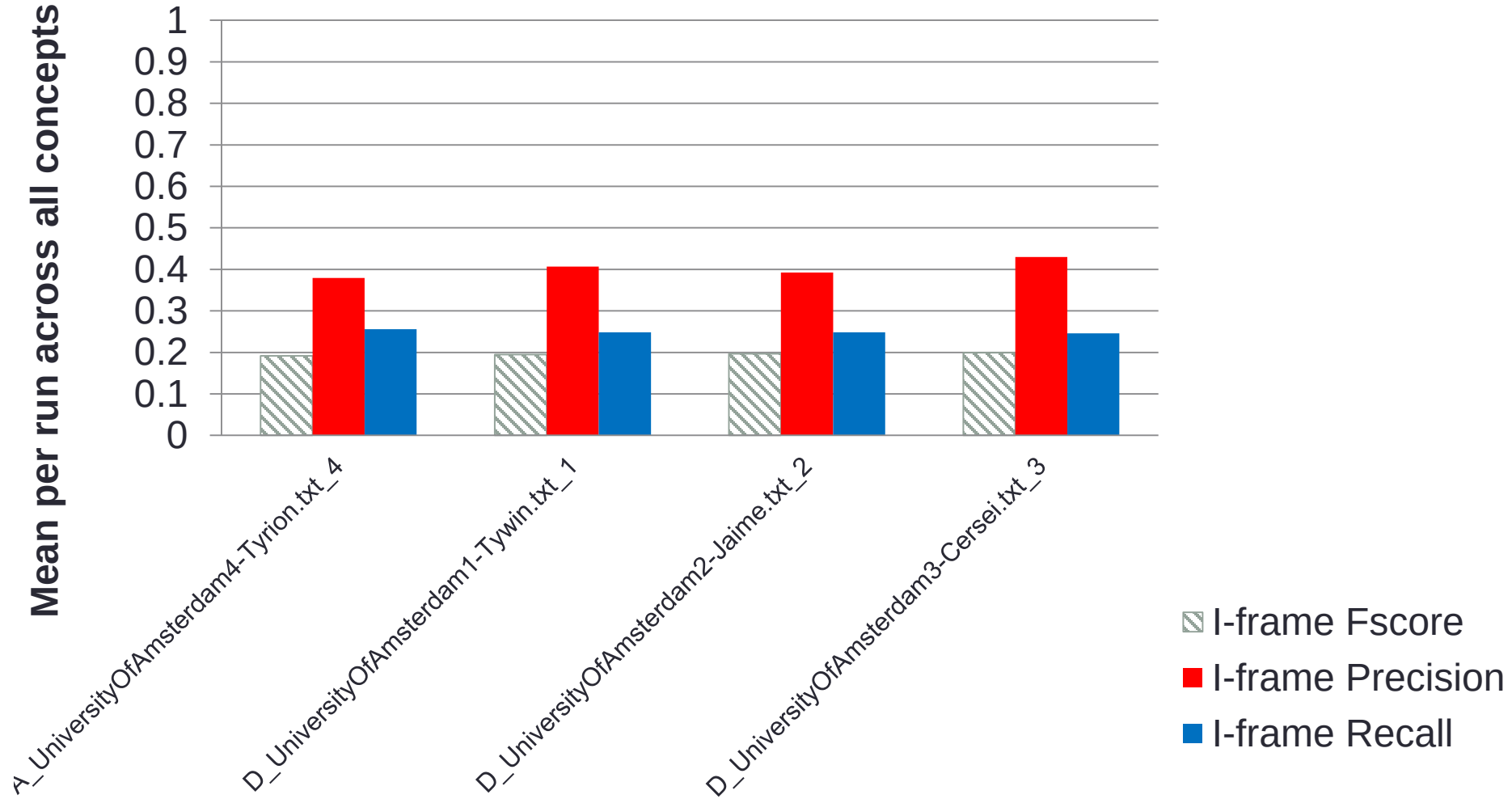
- Temporal localization: precision, recall and fscore based on the judged I-frames.
- Spatial localization: precision, recall and fscore based on the located pixels representing the concept.
- An average of precision, recall and fscore for temporal and spatial localization across all I-frames for each concept and for each run.



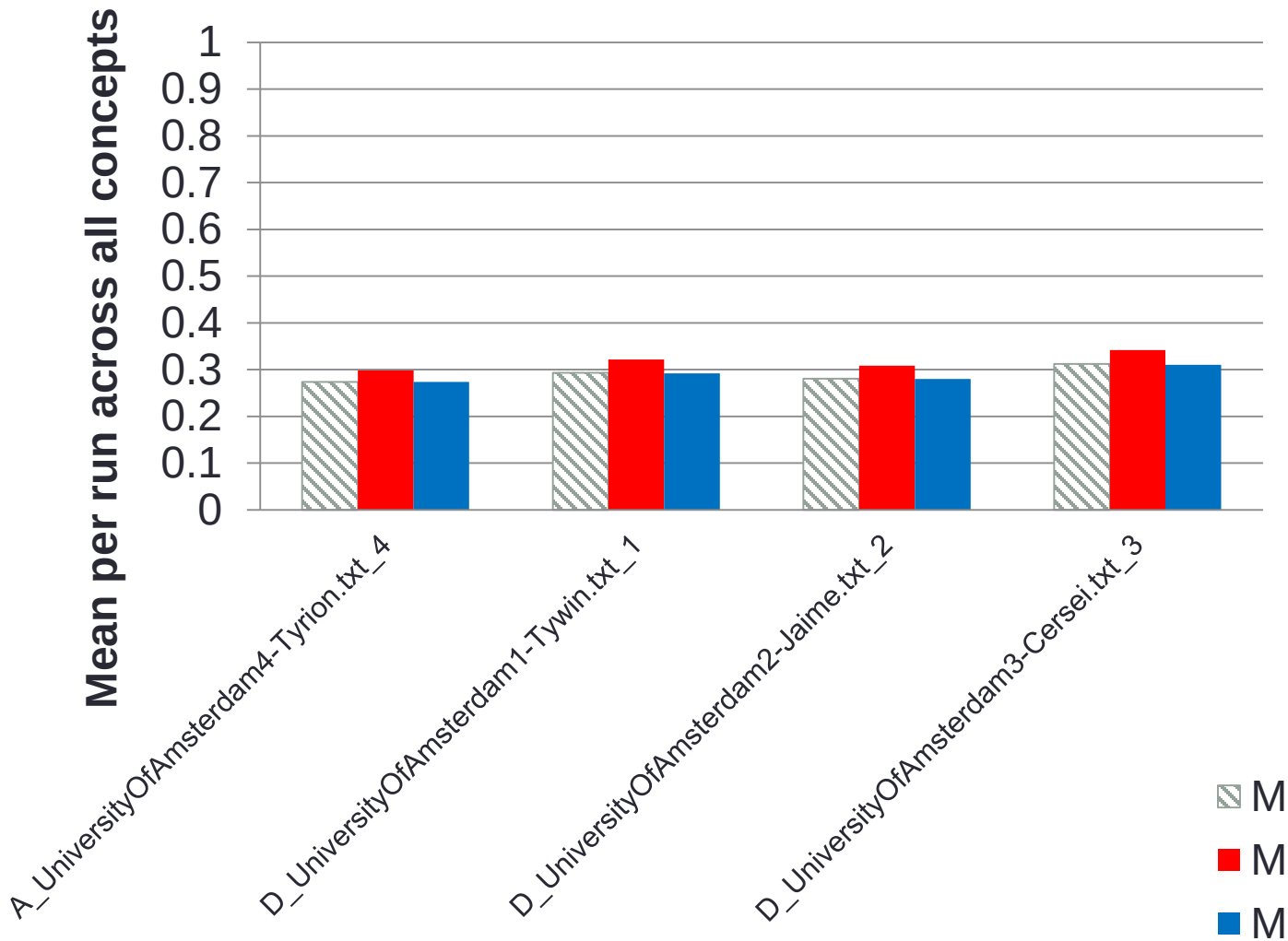
Participants (Finishers)

- 16 teams applied, only 1 team submitted 4 runs!
  - UvA (University Of Amsterdam)

Temporal localization results by run

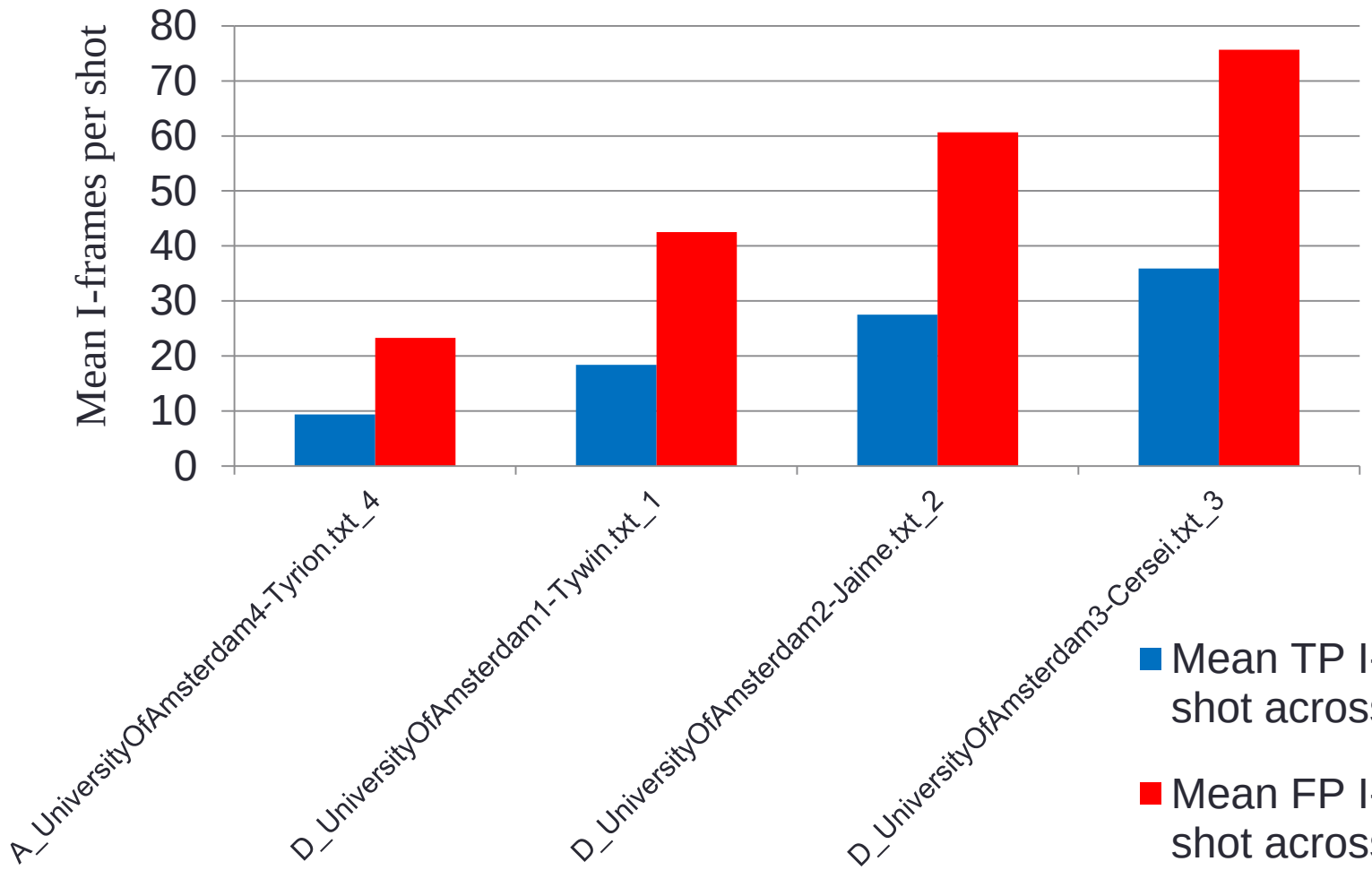


Spatial Localization results by run



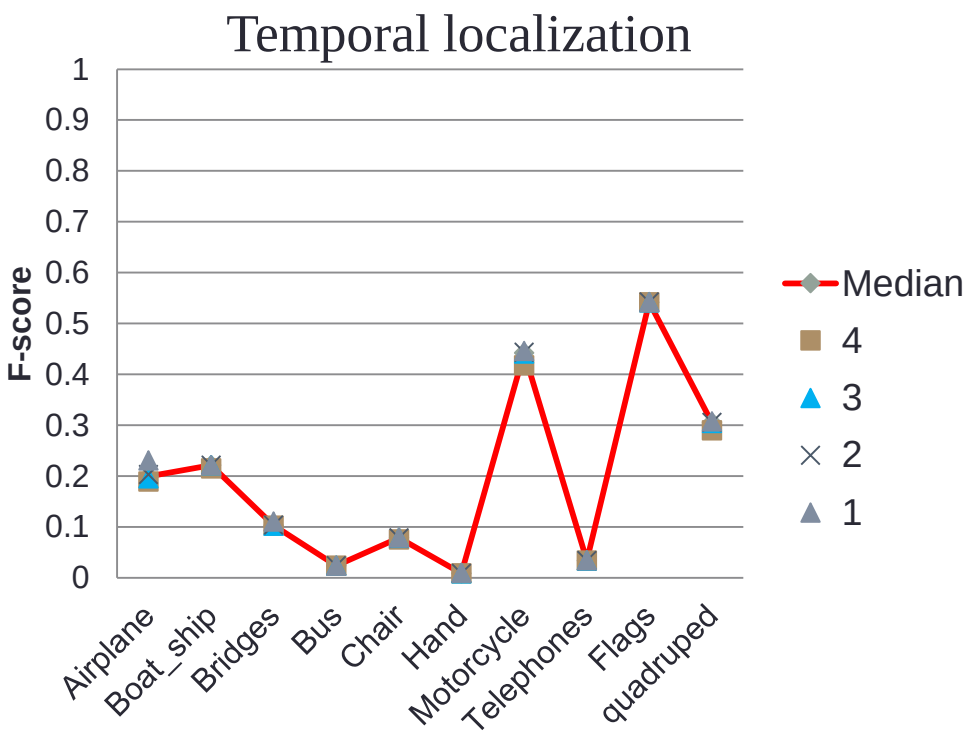
Spatial localization seems to be better than temporal (contrary to 2013 results).
Hard to conclude as all runs come from 1 team

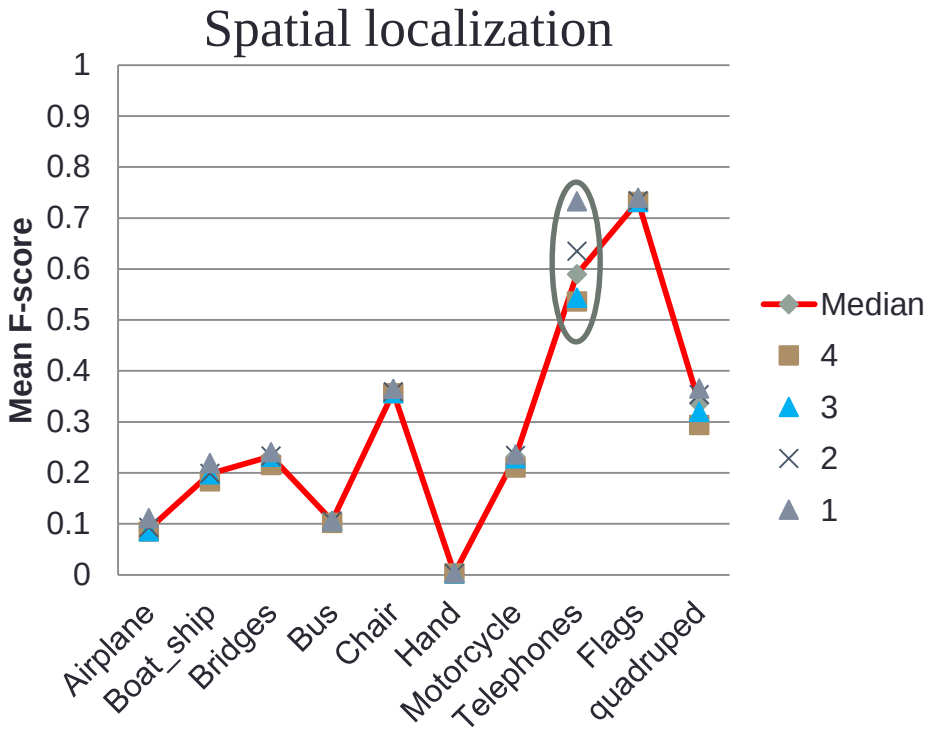
TP vs FP submitted I-frames by run



How can systems find the right balance between TP vs FP I-frames ?

Results per concept





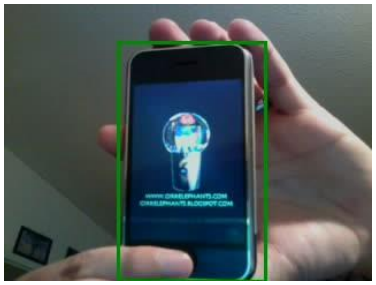
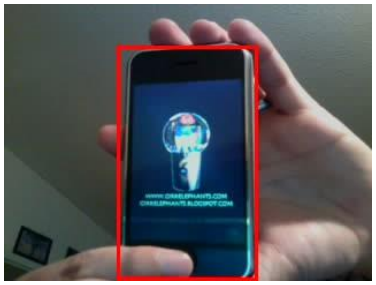
Most concepts are better in spatial localization compared to temporal.
However, 1 team runs are not enough to conclude!

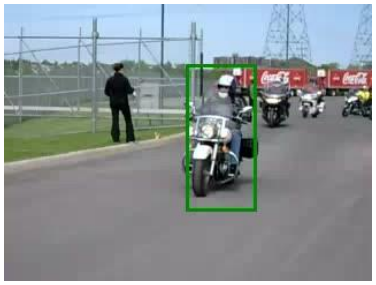


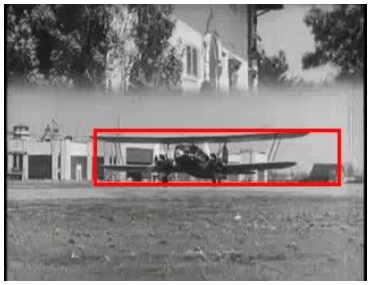
GT
Sys

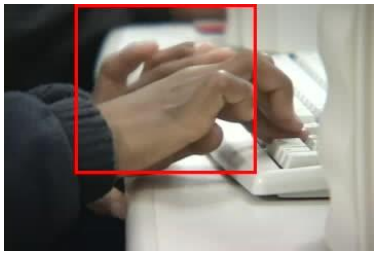
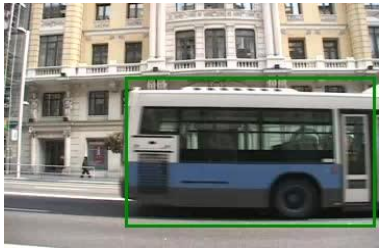




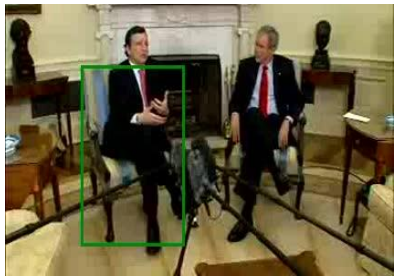



Samples of good localization

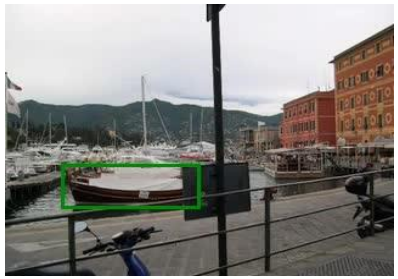
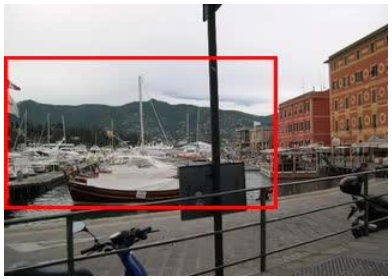


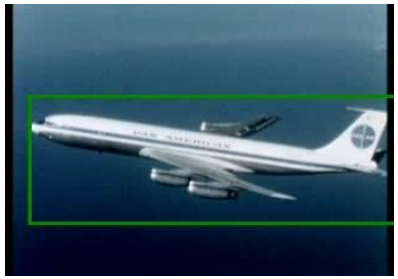
GT
Sys








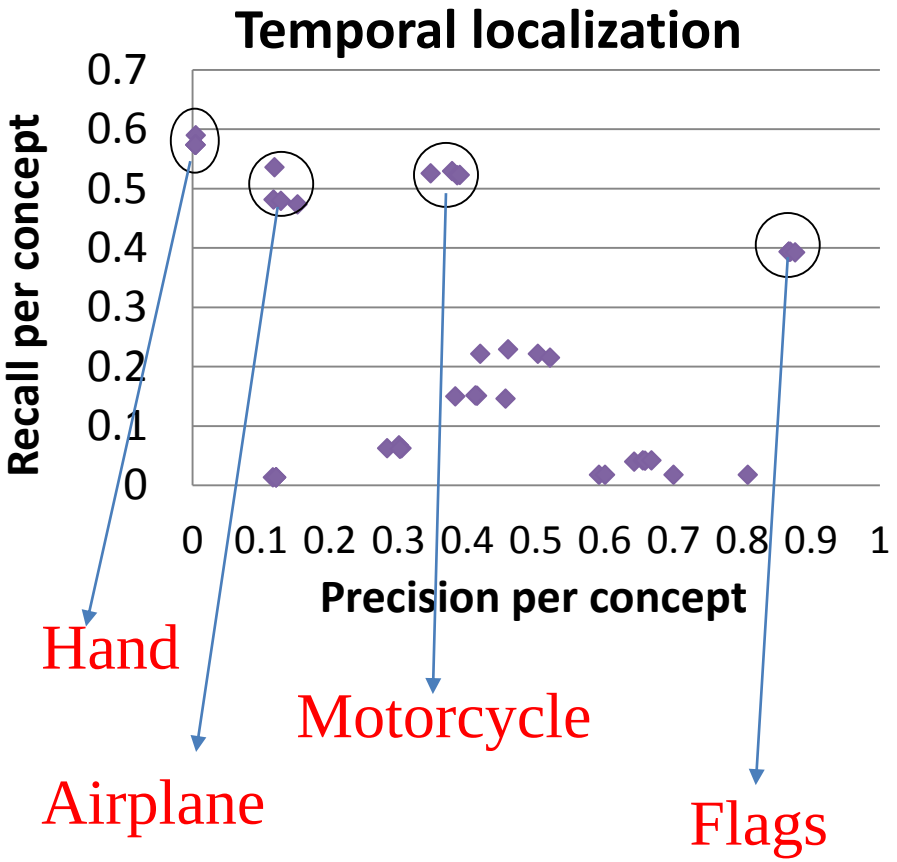



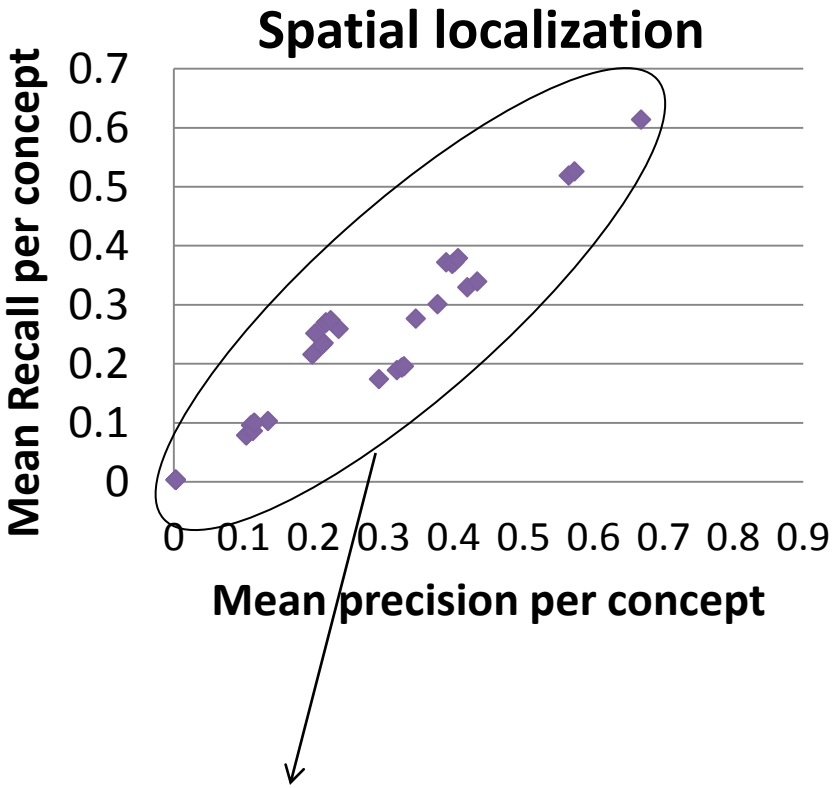






Samples of less good localization

Results per concept across all runs





Submissions missed a lot of TP I-frames in general.

submitted bounding boxes approximate G.T boxes in size with some overlap. Systems are good in finding the real box sizes☺, not so much the real position.

2014 Observations

- 2014 main task was harder than 2013 main task (different data and different set of target concepts)
- Raw system scores have higher Max and Median compared to TV2013, still relatively low.
- Most common concepts with TV2013 have higher median scores.
- Most Progress systems improved significantly from 2013 to 2014.
- Significantly less participants (15 versus 26 for TV2013), most of the loss is in the “long tail”, partly explaining why the median performance is higher even though the task is harder.
- Localization runs missed a lot of TP I-frames.
- Localization submitted boxes approximate true box sizes with some overlap.

2014 Observations

- Approaches similar to TV 2013 with many innovations
- Improved bag of visual words, many dense and pyramidal
- Fisher vectors and similar (VLAD, VLAT, SV...)
- Use of several key frames per shot
- Use of audio features (MFCC+)
- Use of trajectory-based features
- Encoding of spatial information in Fisher vectors
- More semantic features
- Pseudo-relevance feedback
- More deep learning, co-training with ImageNet
- Use of hidden layers in deep convolutional networks
- Fast Local Area Independent Representation for localization
- Hard negative mining

SIN 2015

- Globally keep the task similar and of similar scale, test on IACC.2.C
- Further explore the “no annotation” and “localization” variants
- Sharing of data still proposed by IRIM
- Method for measuring progress over years

  → more progress submission are encouraged

  → we may accept late 2013-2014 progress submissions for a better progress analysis
- Collaborative annotation unchanged
- Feedback welcome

Extra slides for reference

Motivation for xinfAP and pooling strategy

- to make the evaluation more sensitive to shots returned below the lowest rank (~100) previously pooled and judged
- to adjust the sampling to match the relative importance of highest ranked items to average precision
- to exploit more infAP's ability to estimate of AP well even at sampling rates much below the 50% rate used in previous years

NIST median baseline run by NIST

- A median baseline run is created for each run type and training category.
- Basic idea:
  - For each feature, find the median rank of each submitted shot calculated across all submitted runs in that run type and training category.
  - The final shot median rank value is weighted by the ratio of all submitted runs to number of runs that submitted that shot:

$$ShotX_{Median\_rank} = Median\_rank * \frac{TotalNumberOfRuns}{NumberOfRunsSubmittedX}$$

Sharing of data for TRECVID SIN

- Organized by the IRIM groups of CNRS GRD ISIS.
- IRIM proposes its data sharing organization for the TRECVID SIN task. This comprises:
  - a wiki with read-write access for all
  - a data repository with read access for all and currently a write access only via one of the organizers
  - a small set of simple file formats
  - a (quite) simple directory structure
- Shared data mostly consist in descriptors and classification scores.
- Rewarding principle (same as for other contributions)
  - share and be cited and evaluated
  - use freely and cite

Sharing of data for TRECVID SIN

- Wiki (write access with trecvid active participant login/password):
  - http://mrim.imag.fr/trecvid/wiki
  - http://mrim.imag.fr/trecvid/wiki/doku.php?id=sin_2013_task
- Associated data for SIN 2010-2015 (access to some parts with IACC collection login/password):
  - http://mrim.imag.fr/trecvid/sin12
- Related actions:
  - Sharing of low-level descriptors by CMU for TRECVID 2003-2004
  - Mediamill challenge (101 concepts) using TRECVID 2005 data
  - Sharing of detection scores by CU-Vireo on TRECVID 2008-2010 data
- Possible extension to other TRECVID tasks, e.g. MED.
- Currently needs update, announced when finished